Toolbeitrag: DARIAH Topics Explorer

Mareike Schumacher [1]

1. Universität Regensburg

Thema:	Topic Modeling	DOI:	10.48694/fortext.3728
Jahrgang:	1	Ausgabe:	8
Erscheinungsdatum:	2024-07-10	Erstveröffentlichung:	2018-12-10 auf fortext.net
Lizenz:			open access

Allgemeiner Hinweis: Rot dargestellte Begriffe werden im Glossar am Ende des Beitrags erläutert. Alle externen Links sind auch am Ende des Beitrags aufgeführt.

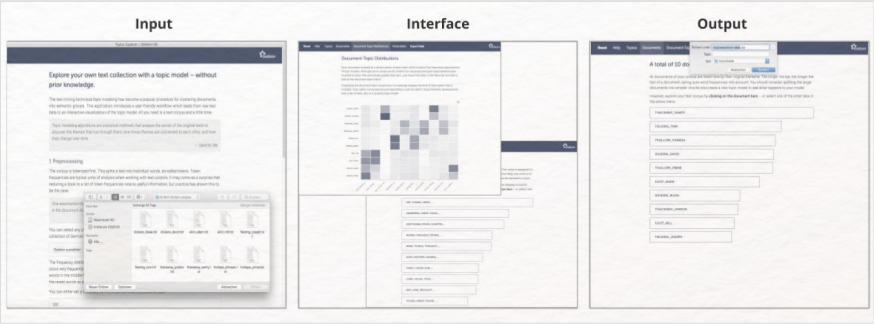

Abb. 1: Der Workflow des DARIAH Topics Explorer: Ein Korpus aus mehreren Texten im TXT- oder XML-Format hochladen und die Parameter festlegen, Daten in Tabellenform und Visualisierungen auf dem Bildschirm ansehen und auswerten, oder eine ZIP-Datei mit Tabellendokumenten und Grafiken als HTML herunterladen.

- **Systemanforderungen:** Der Topics Explorer läuft auf Windows-, Mac- und Linux-PCs, ist offline und ohne Installation nutzbar. *Hinweis*: Mit der Veröffentlichung von macOS 10.15 Catalina hat Apple neue Sicherheitsfunktionen eingeführt, die zu Problemen beim Starten des TopicsExplorer führen. Als vorübergehende Abhilfe folgen Sie den Anweisungen auf dieser Webseite, um die Anwendung aus dem Quellcode zu installieren.
- **Stand der Entwicklung:** Die Version 2.0 des Topics Explorers wurde im Dezember 2018 herausgebracht und wird auch weiterhin fortentwickelt.
- **Herausgeber:** Severin Simmler für DARIAH-DE
- **Lizenz:** Kostenfrei
- **Weblink:** https://github.com/DARIAH-DE/TopicsExplorer
- **Im- und Export:** Der Topics Explorer benötigt mindestens 5 Textdokumente im TXT- (vgl. Reintext-Version) oder XML- Format, um daraus Topics zu generieren. Die Daten und Visualisierungen können in unterschiedlichen Formaten heruntergeladen werden (darunter HTML, Tabellenformate und Bildformate).
- **Sprachen:** Keine Angabe

1. Für welche Fragestellungen kann der DARIAH Topics Explorer eingesetzt werden?

Der Topics Explorer eignet sich besonders zur explorativen Untersuchung (vgl. Distant Reading) von Fragen nach Themenfeldern in größeren Textsammlungen (vgl. Korpus). Das können z. B. wiederkehrende Motive im Werk eines Autors sein oder Themen von Texten einer ganzen Epoche.

2. Welche Funktionalitäten bietet der DARIAH Topics Explorer und wie zuverlässig ist das Tool?

Funktionen:

- Zerlegung von Textdokumenten in Ein-Wort-Listen (Tokenization (vgl. Type/Token))
- Verwendung von Stoppwortlisten
- Auswahl einer Anzahl von Topics
- Auswahl einer Anzahl von Durchgängen des Machine Learning-Prozesses

- Visualisierung der Topics

Zuverlässigkeit: Der Topics Explorer wurde für DH-Einsteiger*innen entwickelt, die Topic Modeling mit Hilfe einer grafischen Nutzeroberfläche (vgl. GUI) durchführen möchten. Das Tool ist nicht darauf ausgelegt, große Textsammlungen auszuwerten. Mit kleineren Sammlungen (bis zu 100 oder 200 Texte) läuft das Tool sehr zuverlässig. Die Erstellung der Topics läuft vergleichsweise zügig, kann jedoch nach Größe der Textsammlung variieren.

3. Ist der DARIAH Topics Explorer für DH-Einsteiger*innen geeignet?

Checkliste	√ / teilweise /-
Methodische Nähe zur traditionellen Literaturwissenschaft	-
Grafische Benutzeroberfläche	√
Intuitive Bedienbarkeit	√
Leichter Einstieg	√
Handbuch vorhanden	-
Handbuch aktuell	-
Tutorials vorhanden	-
Erklärung von Fachbegriffen	-
Gibt es eine gute Nutzerbetreuung?	teilweise

Der Topics Explorer setzt zwar keine technischen Fertigkeiten, dafür aber Grundkenntnisse in der Methodik des Topic Modeling voraus (vgl. Topic Modeling (Horstmann 2024)). Fachbegriffe wie LDA werden genannt, aber nicht weiter erklärt. Alle erklärenden Texte sind in englischer Sprache verfasst. Die Nutzerbetreuung findet über GitHub statt und eine Problemlösung kann nach eigenen Angaben der Entwickler unterschiedlich viel Zeit in Anspruch nehmen.

4. Wie etabliert ist der DARIAH Topics Explorer in den (Literatur-)Wissenschaften?

Der Topics Explorer wurde im Frühjahr 2018 als Beta-Version und im Dezember desselben Jahres als Version 2.0 herausgegeben. Damit ist das Tool zum jetzigen Zeitpunkt noch sehr jung und wurde bisher nicht in literaturwissenschaftlichen Publikationen erwähnt.

5. Unterstützt der DARIAH Topics Explorer kollaboratives Arbeiten?

Nein, der Topics Explorer hat keine Funktionalitäten, die kollaborativ genutzt werden können.

6. Sind meine Daten beim DARIAH Topics Explorer sicher?

Ja. Für die Nutzung des Topics Explorers ist keine Angabe persönlicher Daten notwendig. Die verarbeiteten Textdaten bleiben auf dem eigenen PC. Die Nutzung des Topics Explorers ist also unter persönlichkeitsrechtlichen und auch unter urheberrechtlichen Gesichtspunkten unproblematisch.

Externe und weiterführende Links

- DARIAH Topics Explorer auf GitHub: https://dariah-de.github.io/TopicsExplorer/#the-application (Letzter Zugriff: 17.09.2024)
- DARIAH TOpics Exlporer Nutzer*innenbetreuung: https://web.archive.org/save/https://github.com/DARIAH-DE/TopicsExplorer/issues (Letzter Zugriff: 17.09.2024)

Bibliographie

Horstmann, Jan. 2024. Methodenbeitrag: Topic Modeling. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 8. Topic Modeling (7. Oktober). doi: 10.48694/fortext.3717, https://fortext.net/routinen/methoden/topic-modeling.

Glossar

Annotation Annotation beschreibt die manuelle oder automatische Hinzufügung von Zusatzinformationen zu einem Text. Die manuelle Annotation wird händisch durchgeführt, während die (teil-)automatisierte Annotation durch Machine-Learning-Verfahren durchgeführt wird. Ein klassisches Beispiel ist das automatisierte PoS-Tagging (Part-of-Speech-Tagging), welches oftmals als Grundlage (Preprocessing) für weitere Analysen wie Named Entity Recognition (NER) nötig ist. Annotationen können zudem deskriptiv oder analytisch sein.

Browser Mit Browser ist in der Regel ein Webbrowser gemeint, also ein Computerprogramm, mit dem das Anschauen, Navigieren auf, und Interagieren mit Webseiten möglich wird. Am häufigsten genutzt werden dafür Chrome, Firefox, Safari oder der Internet Explorer.

Close Reading Close Reading bezeichnet die sorgfältige Lektüre und Interpretation eines einzelnen oder weniger Texte. Close Reading ist in der digitalen Literaturwissenschaft außerdem mit der manuellen Annotation textueller Phänomene verbunden (vgl. auch Distant Reading als Gegenbegriff).

Commandline Die Commandline (engl. *command line interface* (CLI)), auch Kommandozeile, Konsole, Terminal oder Eingabeaufforderung genannt, ist die direkteste Methode zur Interaktion eines Menschen mit einem Computer. Programme ohne eine grafische Benutzeroberfläche (GUI) werden i. d. R. durch Texteingabe in die Commandline gesteuert. Um die Commandline zu öffnen, klicken Sie auf Ihrem Mac „cmd“ + „space“, geben „Terminal“ ein und doppelklicken auf das Suchergebnis. Bei Windows klicken Sie die Windowstaste + „R“, geben „cmd.exe“ ein und klicken Enter.

CSV CSV ist die englische Abkürzung für *Comma Separated Values*. Es handelt sich um ein Dateiformat zur einheitlichen Darstellung und Speicherung von einfach strukturierten Daten mit dem Kürzel `.csv`, sodass diese problemlos zwischen IT-Systemen ausgetauscht werden können. Dabei sind alle Daten zeilenweise angeordnet. Alle Zeilen wiederum sind in einzelne Datenfelder aufgeteilt, welche durch Trennzeichen wie Semikola oder Kommata getrennt werden können. In Programmen wie Excel können solche Textdateien als Tabelle angezeigt werden.

Distant Reading Distant Reading ist ein Ansatz aus den digitalen Literaturwissenschaften, bei dem computationelle Verfahren auf häufig große Mengen an Textdaten angewandt werden, ohne dass die Texte selber gelesen werden. Meist stehen hier quantitative Analysen im Vordergrund, es lassen sich jedoch auch qualitative Metadaten quantitativ vergleichen. Als Gegenbegriff zu *Close Reading* wurde der Begriff insbesondere von Franco Moretti (2000) geprägt.

GUI GUI steht für *Graphical User Interface* und bezeichnet eine grafische Benutzeroberfläche. Ein GUI ermöglicht es, Tools mithilfe von grafischen Schaltflächen zu bedienen, um somit beispielsweise den Umgang mit der Commandline zu umgehen.

HTML HTML steht für *Hypertext Markup Language* und ist eine textbasierte Auszeichnungssprache zur Strukturierung elektronischer Dokumente. HTML-Dokumente werden von Webbrowsern dargestellt und geben die Struktur und Online-Darstellung eines Textes vor. HTML-Dateien können außerdem zusätzliche Metainformationen enthalten, die auf einer Webseite selbst nicht ersichtlich sind.

Korpus Ein Textkorpus ist eine Sammlung von Texten. Korpora (Plural für „das Korpus“) sind typischerweise nach Textsorte, Epoche, Sprache oder Autor*in zusammengestellt.

LDA LDA steht für *Latent Dirichlet Allocation* und ist ein generatives, statistisches Wahrscheinlichkeitsmodell, welches zum Topic Modeling angewendet werden kann. Bei der LDA werden auf Grundlage eines Wahrscheinlichkeitsmodells Wortgruppen aus Textdokumenten erstellt. Dabei wird jedes Dokument als eine Mischung von verborgenen Themen betrachtet und jedes Wort einem Thema zugeordnet. Wortreihenfolgen und Satzzusammenhänge spielen dabei keine Rolle.

Lemmatisieren Die Lemmatisierung von Textdaten gehört zu den wichtigen Preprocessing-Schritten in der Textverarbeitung. Dabei werden alle Wörter (Token) eines Textes auf ihre Grundform zurückgeführt. So werden beispielsweise Flexionsformen wie „schneller“ und „schnelle“ dem Lemma „schnell“ zugeordnet.

Machine Learning Machine Learning, bzw. maschinelles Lernen im Deutschen, ist ein Teilbereich der künstlichen Intelligenz. Auf Grundlage möglichst vieler (Text-)Daten erkennt und erlernt ein Computer die häufig sehr komplexen Muster und Gesetzmäßigkeiten bestimmter Phänomene. Daraufhin können die aus den Daten gewonnen Erkenntnisse verallgemeinert werden und für neue Problemlösungen oder für die Analyse von bisher unbekannten Daten verwendet werden.

Markup (Textauszeichung) Die Textauszeichnung (eng. *Markup*) fällt in den Bereich der Daten- bzw. Textverarbeitung, genauer in das Gebiet der Textformatierung, welche durch Auszeichnungssprachen wie XML implementiert wird. Dabei geht es um die Beschreibung, wie einzelne Elemente eines Textes beispielsweise auf Webseiten grafisch dargestellt werden sollen.

Markup Language Markup Language bezeichnet eine maschinenlesbare Auszeichnungssprache, wie z. B. HTML, zur Formatierung und Gliederung von Texten und anderen Daten. So werden beispielsweise

auch Annotationen durch ihre Digitalisierung oder ihre digitale Erstellung zu Markup, indem sie den Inhalt eines Dokumentes strukturieren.

Metadaten Metadaten oder Metainformationen sind strukturierte Daten, die andere Daten beschreiben. Dabei kann zwischen administrativen (z. B. Zugriffsrechte, Lizenzierung), deskriptiven (z. B. Textsorte), strukturellen (z. B. Absätze oder Kapitel eines Textes) und technischen (z. B. digitale Auflösung, Material) Metadaten unterschieden werden. Auch Annotationen bzw. Markup sind Metadaten, da sie Daten/Informationen sind, die den eigentlichen Textdaten hinzugefügt werden und Informationen über die Merkmale der beschriebenen Daten liefern.

Named Entities Eine Named Entity (NE) ist eine Entität, oft ein Eigenname, die meist in Form einer Nominalphrase zu identifizieren ist. Named Entities können beispielsweise Personen wie „Nils Holgerson“, Organisationen wie „WHO“ oder Orte wie „New York“ sein. Named Entities können durch das Verfahren der Named Entity Recognition (NER) automatisiert ermittelt werden.

POS PoS steht für *Part of Speech*, oder „Wortart“ auf Deutsch. Das PoS- Tagging beschreibt die (automatische) Erfassung und Kennzeichnung von Wortarten in einem Text und ist of ein wichtiger Preprocessing-Schritt, beispielsweise für die Analyse von Named Entities.

Preprocessing Für viele digitale Methoden müssen die zu analysierenden Texte vorab „bereinigt“ oder „vorbereitet“ werden. Für statistische Zwecke werden Texte bspw. häufig in gleich große Segmente unterteilt (*chunking*), Großbuchstaben werden in Kleinbuchstaben verwandelt oder Wörter werden lemmatisiert.

Reintext-Version Die Reintext-Version ist die Version eines digitalen Textes oder einer Tabelle, in der keinerlei Formatierungen (Kursivierung, Metadatenauszeichnung etc.) enthalten sind. Reintext-Formate sind beispielsweise TXT, RTF und CSV.

Stoppwortliste Stoppwörter sind hochfrequente Wörter, meist Funktionswörter, die, aufgrund ihrer grammatisch bedingten Häufigkeit, beispielsweise die Ergebnisse von inhaltlichen oder thematischen Analysen verzerren können. Deshalb werden diese Wörter, gesammelt in einer Stoppwortliste, bei digitalen Textanalysen meist nicht berücksichtigt.

TEI Die *Text Encoding Initiative* (TEI) ist ein Konsortium, das gemeinsam einen Standard für die Darstellung von Texten in digitaler Form entwickelt. Die TEI bietet beispielsweise Standards zur Kodierung von gedruckten Werken und zur Auszeichnung von sprachlichen Informationen in maschinenlesbaren Texten (siehe auch XML und Markup).

Topic Modeling Das Topic Modeling ist ein statistisches, auf Wahrscheinlichkeitsrechnung basierendes, Verfahren zur thematischen Exploration größerer Textsammlungen. Das Verfahren erzeugt „Topics“ zur Abbildung häufig gemeinsam vorkommender Wörter in einem Text. Für die Durchführung können verschiedene Algorithmen und Modelle wie das LDA verwendet werden.

Type/Token Das Begriffspaar „Type/Token“ wird grundsätzlich zur Unterscheidung von einzelnen Vorkommnissen (Token) und Typen (Types) von Wörtern oder Äußerungen in Texten genutzt. Ein Token ist also ein konkretes Exemplar eines bestimmten Typs, während ein Typ eine im Prinzip unbegrenzte Menge von Exemplaren (Token) umfasst.
Es gibt allerdings etwas divergierende Definitionen zur Type-Token-Unterscheidung. Eine präzise Definition ist daher immer erstrebenswert. Der Satz „Ein Bär ist ein Bär.“ beinhaltet beispielsweise fünf Worttoken („Ein“, „Bär“, „ist“, „ein“, „Bär“) und drei Types, nämlich: „ein“, „Bär“, „ist“. Allerdings könnten auch vier Types, „Ein“, „ein“, „Bär“ und „ist“, als solche identifiziert werden, wenn Großbuchstaben beachtet werden.

XML XML steht für *Extensible Markup Language* und ist eine Form von Markup Language, die sowohl computer- als auch menschenlesbar und hochgradig anpassbar ist. Dabei werden Textdateien hierarchisch strukturiert dargestellt und Zusatzinformationen i. d. R. in einer anderen Farbe als der eigentliche (schwarz gedruckte) Text dargestellt. Eine standardisierte Form von XML ist das TEI-XML.